W628–W632 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 11 June 2012*
doi:10.1093/nar/gks550

BiForce Toolbox: powerful high-throughput computational analysis of gene–gene interactions in genome-wide association studies

Attila Gyenesei[1], Jonathan Moody[2], Asta Laiho[1], Colin A.M. Semple[2], Chris S. Haley[2] and Wen-Hua Wei[2,*]

[1]Finnish Microarray and Sequencing Centre, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, 20520, Turku, Finland and [2]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine at the University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, UK

Received February 26, 2012; Revised May 11, 2012; Accepted May 16, 2012

ABSTRACT

Genome-wide association studies (GWAS) have discovered many loci associated with common disease and quantitative traits. However, most GWAS have not studied the gene–gene interactions (epistasis) that could be important in complex trait genetics. A major challenge in analysing epistasis in GWAS is the enormous computational demands of analysing billions of SNP combinations. Several methods have been developed recently to address this, some using computers equipped with particular graphical processing units, most restricted to binary disease traits and all poorly suited to general usage on the most widely used operating systems. We have developed the BiForce Toolbox to address the demand for high-throughput analysis of pairwise epistasis in GWAS of quantitative and disease traits across all commonly used computer systems. BiForce Toolbox is a stand-alone Java program that integrates bitwise computing with multithreaded parallelization and thus allows rapid full pairwise genome scans via a graphical user interface or the command line. Furthermore, BiForce Toolbox incorporates additional tests of interactions involving SNPs with significant marginal effects, potentially increasing the power of detection of epistasis. BiForce Toolbox is easy to use and has been applied in multiple studies of epistasis in large GWAS data sets, identifying interesting interaction signals and pathways.

INTRODUCTION

Genome-wide association studies (GWAS) have successfully discovered nearly 1500 genetic loci associated with over 200 common disease and quantitative traits (http://www.genome.gov/GWAStudies/). However, most GWAS have been unable to study the gene–gene interactions (epistasis) that could be a potential source of 'missing heritability' evidenced in these studies (1,2). A recent study demonstrated that epistasis could exist widely in biological pathways and create substantial phantom heritability undetectable via conventional GWAS (3). Indeed, additional tests of interactions involving GWAS loci with significant marginal effects have successfully discovered epistasis in several studies (4–7). However, despite enormous effort, so far studying epistasis in GWAS has been far more challenging and less fruitful than conventional GWAS focussing on single locus effects (3,8).

One major challenge in studying epistasis in GWAS (typically with hundreds of thousands of SNP markers) is the need to scrutinize billions of pairwise SNP combinations in order to consider all possible interactions. The challenge is increasing as more sequencing data (and consequently more SNPs) become available in future GWAS. Several methods have been developed recently to tackle this challenge (5,9–13). These methods are confined to either binary disease or quantitative traits and several are designed specifically for computers equipped with particular graphical processing units. Substantial computing time is still required in some of these methods, especially when widely available computer systems are used (11,13). There is therefore great demand for high-throughput tools that can run on widely used computer platforms to analyse both quantitative and disease traits, making the analysis of epistasis routine in GWAS and ultimately

*To whom correspondence should be addressed. Tel: +44 131 332 2471; Fax: +44 131 467 8456; Email: wenhua.wei@igmm.ed.ac.uk

improving our understanding of the role of epistasis in complex traits.

Another major challenge is that many current GWAS populations may have only limited power to detect and replicate significant epistasis signals due to their relatively small sample sizes (8,14). The approach of meta-analysis of multiple GWAS (15) may enhance the power of detection of epistasis but this is not yet applicable due to the lack of powerful computational tools to support epistasis analysis in such data sets using imputed (not categorical) genotype data. Pathway-based approaches may narrow the search space and also enhance power, for example, by seeking pathway–pathway interactions (3) or by identifying common pathways enriched in epistatic genes with modest interaction signals (i.e. strong but not necessarily genome-wide significant) detected from multiple GWAS populations (16). In either case, fast screening of pairwise interactions in individual GWAS populations appears to be critical in order to provide information for pathway-based analyses.

We have developed the BiForce Toolbox to address the challenges of high-throughput detection of epistasis. BiForce Toolbox is programmed in Java to support large scale analysis of pairwise epistasis in quantitative and disease traits on commonly used computer systems (e.g. running either Windows, OSX or Linux operating system) via a graphical user interface (GUI) or the command line. It integrates available algorithms and advanced computing technologies such as bitwise computing technologies first adopted in BOOST (12) and FastEpistasis (11) and multi-threaded parallelization into one software package to offer rapid and comprehensive pairwise genome scans. BiForce Toolbox is built on concepts and algorithms we previously developed to maximize the power of detection of different forms of epistasis while controlling false positive rates (9,17,18). It has been rigorously tested and successfully identified interesting interaction signals and pathways in multiple studies using large GWAS data sets (7,16). Here, we describe the main features and functionality of the BiForce Toolbox using the body mass index (BMI) trait in the Northern Finland Birth Cohort 1966 (NFBC1966) (16,19) as an example. The NFBC1966 GWAS data were provided by the database of Genotype and Phenotype (dbGaP; http://www.ncbi.nlm.nih.gov/gap) via specific Data Use Certification.

PROGRAM OVERVIEW

BiForce Toolbox is named to reflect its main features: fast screening of pairwise interactions in GWAS of complex disease and quantitative traits, using the brute force computational power of bitwise computing and multi-threaded parallelization. It is implemented in Java to enable its use on most commonly used computer systems, allowing local secure analysis of GWAS data sets and comprehensive fast genome-wide scans for epistasis. It has been designed to be user friendly and benefits from an intuitive GUI as well as command line access for automated submission of jobs. BiForce Toolbox uses a combined search algorithm (17,18) that integrates a full pairwise genome scans with specific tests of interactions involving SNPs with marginal effects that are genome-wide significant (marginal-SNPs) to increase the power of detection. BiForce Toolbox is free and the binary files compiled for the three operating systems (Mac OSX, Windows and Linux) can be downloaded from http://bioinfo.utu.fi/BiForceToolbox/.

BiForce Toolbox takes ordinary GWAS data (in genotype and phenotype files) as input, where SNP genotypes need to pass normal quality control procedures and phenotypes are recommended to be adjusted for covariates and relatedness for quantitative traits (7,16) or post BiForce analysis for disease traits. After data input, it converts the SNP genotype data into Boolean bit values and the data are stored in memory-efficient Java BitSet arrays that allow missing SNP genotypes to be handled easily and Boolean bitwise operations (e.g. logical AND) to be applied to the arrays of bit values, which makes the association tests (see below) extremely fast. Further, BiForce Toolbox partitions the whole pairwise genome scan automatically into smaller tasks and feed them evenly to available processing threads to compute in parallel and store results appropriately.

The combined search algorithm implemented in BiForce Toolbox includes two consecutive genome scans: single SNP-based genome-wide association tests and pairwise epistatic interaction tests of all SNP combinations. Marginal-SNPs are identified in the first scan and used to test interactions involving them. By default the 5% genome-wide significance thresholds are derived based on the Bonferroni correction for the total number of tests performed. Given N to be the total number of SNPs in a GWAS with K ($K > 0$) marginal-SNPs being identified, the thresholds are: $P = 0.05/N$ for marginal-SNPs, $P = 0.05/[(N-1) * K]$ for marginal-SNP interactions and $P = 0.05/[N * (N-1)/2]$ for a pairwise genome scan. Alternatively, user specified significance thresholds can be specified in the analysis. The algorithm currently concerns SNPs on the autosomal chromosomes only.

Association tests are based on linear regression models, where the genotypes of each SNP (i.e. homozygote of the minor allele, homozygote of the major allele and heterozygote) are fitted as fixed factors. Single SNP-based association tests are straightforward each with two degrees of freedom. Pairwise SNP interactions are assessed using contingency tables which makes BiForce Toolbox applicable to both quantitative and binary disease traits. Briefly, a saturated model (fitting two SNPs and interactions) is tested against a reduced model (fitting the two SNPs without interactions) with four degrees of freedom and then the F ratio for quantitative traits or log-likelihood ratio for binary traits is calculated. P values are then derived according to specific test statistic distributions and appropriate degrees of freedom (assuming fixed four degrees of freedom in interaction tests for disease traits). For disease traits, BiForce Toolbox adopts the approximation filtering algorithm developed in BOOST (12) as a default option to accelerate the exhaustive pairwise genome scan which can be dismissed when necessary (i.e. using only log-likelihood ratio tests in the scan).

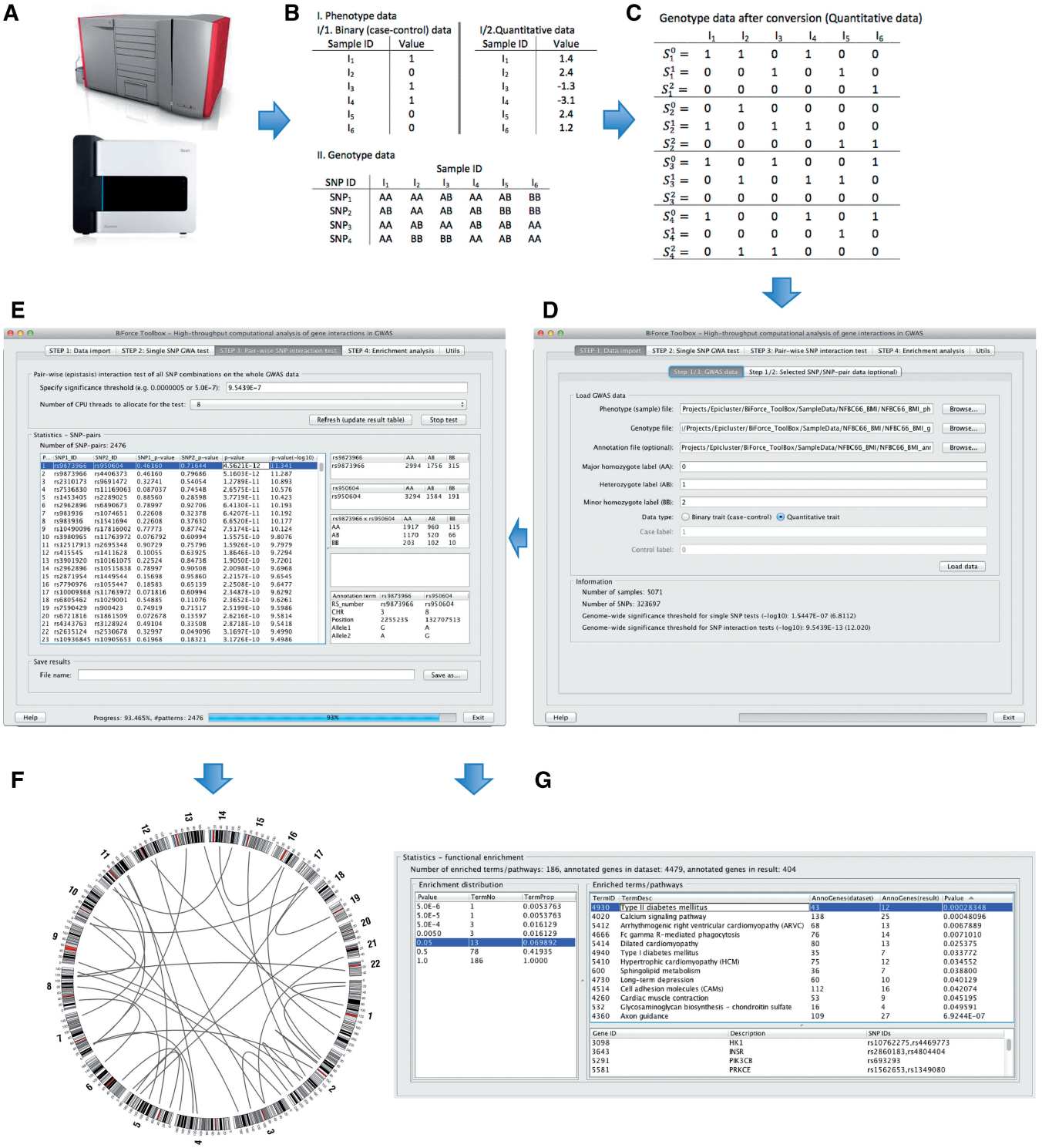

Figure 1. BiForce Toolbox workflow. GWAS data are generated by microarray or next-generation sequencing platforms (**A**). Example of input data formats for disease and quantitative traits (**B**). Example of genotype data conversion and storage in bitwise arrays (**C**). Snapshot of the BiForce Toolbox GUI after loading and conversion of the BMI data of the NFBC1966 cohort (note the multi-tab options such as working with subsets of SNPs or SNP-pairs and performing single SNP-based genome-wide association) (**D**). Snapshot of the BiForce Toolbox GUI running the pairwise genome scan of the BMI data with progress reporting at the bottom (note SNP information, genotype counts, contingency table are showed for a retained pair of SNPs; the retained results of the pairwise genome scan can be exported as tab-separated text file or Excel compatible spreadsheet for further analyses by users) (**E**). A graphic view of positions of interaction signals in BMI of the NFBC1966 cohort generated by a third party tool Circos, where chromosome ideograms are shown around the outer ring and are oriented pter–qter in a clockwise direction with centromeres indicated as red bands (**F**). Snapshot of the BiForce Toolbox GUI running pathway enrichment analysis after the pairwise genome scan using epistatic genes with interaction $-\log_{10} P$ values greater than 7.3 (i.e. $P < 5.0e^{-08}$; note the member epistatic genes and associated SNPs are displayed for a pathway selected) (**G**).

Table 1. A short list of the results of the pairwise genome scan of BMI in NFBC1966 cohort[a]

SNP1	SNP2	Anova_Pair	Anova_Int	Pair_$-\log_{10}$P	Int_$-\log_{10}$P	GenoClassNo
rs9873966	rs950604	7.704	14.850	9.587	11.341	9
rs9873966	rs4406373	7.646	14.785	9.496	11.287	9
rs2310173	rs9691472	7.615	14.311	9.448	10.893	9
rs7536830	rs11169063	7.644	13.928	9.492	10.576	9
rs1453405	rs2289025	7.237	13.744	8.862	10.423	9
rs2962896	rs6890673	6.810	13.466	8.203	10.193	9
rs983936	rs1074651	7.380	13.465	9.083	10.192	9
rs983936	rs1541694	7.324	13.447	8.997	10.177	9
rs10490096	rs17816002	7.305	16.789	8.018	10.124	8
rs3980965	rs11763972	7.260	13.001	8.897	9.808	9

[a]SNP1 (SNP2): the first (second) SNP; Anova_Pair (Anova_Int): F ratio of a whole pair of SNPs with interaction (interaction between a pair of SNPs); Pair_$-\log_{10}$P (Int_$-\log_{10}$P): $-\log_{10}$ *P* value of a whole pair of SNPs with interaction (interaction between a pair of SNPs); GenoClassNo: the number of joint genotype classes (9 in total) with samples.

On completion of the search process, BiForce Toolbox can be used to generate summary information and examine the identified SNP pairs (e.g. allele frequencies and contingency tables) after reloading the input data files and the retained results (Figure 1). Furthermore, it can perform a generic analysis of pathways enriched within groups of genes showing interaction signals to provide an initial view of the potential biology underlying these signals. Such an analysis involves two steps: annotating SNPs (with reference SNP ID numbers) to the nearest genes and then performing overrepresentation analysis based on hypergeometric tests against a variety of gene-centred functional data such as Gene Ontology (GO) and KEGG pathway annotation (20). Further details of such an analysis can be found in the manual of BiForce Toolbox.

A TYPICAL ANALYSIS PROCEDURE

We use BMI of the NFBC1966 cohort to illustrate a typical analysis procedure using BiForce Toolbox. The genotype and phenotype data were pre-processed following the instructions given in the original GWAS (19). BMI was further corrected for covariates and relatedness and normalized to prevent spurious associations as detailed elsewhere (16). In total 323 697 autosomal SNPs and 5071 individuals entered BiForce Toolbox for analysis. No marginal-SNPs were identified in the single SNP-based genome scan. The remaining analysis is summarized in Figure 1, with an example list of the retained results of the pairwise genome scan (none is genome-wide significant) shown in Table 1.

PERFORMANCE PROFILE

We measured the BiForce Toolbox performance in analysing disease (50% cases and 50% controls) and quantitative traits of GWAS datasets with 500 000 SNPs and 5000 samples, using one thread and eight threads on a single workstation (2.8 GHz Intel Core iMac with 4 GB RAM, 4 CPU cores each with 2 threads) and 256 threads on a computer cluster (32 nodes each with 4 CPU cores each with 2 threads), respectively. BiForce Toolbox took 118.18, 30.8 and 0.46 h, respectively, for the disease trait, and 293.24 (using 8 threads on the workstation) and 6.81 h (using 256 threads on the cluster) for the quantitative trait. In contrast, FastEpistasis, a parallel extension of PLINK (21) took 29, 4 or 0.5 days to analyse a quantitative trait in a GWAS data set of the same size using 8, 64 or 512 MPI-bound processors, respectively (11); GBOOST, a graphical processing unit version of BOOST (12) took 1.34 h to analyse a disease trait in a smaller GWAS data set (351 542 SNPs and 5003 samples) on a computer with a Nvidia GeForce GTX 285 display card (i.e. 240 CPU cores) (13).

CONCLUSIONS

BiForce Toolbox is a powerful and accessible tool to support high-throughput analysis of epistasis in GWAS of disease and quantitative traits on general computer platforms. It is hoped that with BiForce Toolbox analysis the study of epistasis in GWAS will become a routine exercise and hence improve our understanding of the role of epistasis in the architecture of complex traits.

ACKNOWLEDGEMENTS

We thank the editor and three anonymous reviewers for their valuable suggestions and comments. We are grateful for assistance from the authors of the BOOST software. We acknowledge data access to NHLBI STAMPEED study (Northern Finland Birth Cohort 1966, phs000276.v1.p1) via dbGaP (http://www.ncbi.nlm.nih.gov/gap).

FUNDING

The Biotechnology and Biological Sciences Research Council (BBSRC) [BB/H024484/1 to W.H.W.]; the Medical Research Council Core Fund (to C.A.S., C.S.H. and W.H.W.). Funding for open access charge: BBSRC [BB/H024484/1].

Conflict of interest statement. None declared.

REFERENCES

1. Eichler,E.E., Flint,J., Gibson,G., Kong,A., Leal,S.M., Moore,J.H. and Nadeau,J.H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.
2. Gibson,G. (2010) Hints of hidden heritability in GWAS. *Nat. Genet.*, **42**, 558–560.
3. Zuk,O., Hechter,E., Sunyaev,S.R. and Lander,E.S. (2012) The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA*, **109**, 1193–1198.
4. Evans,D.M., Spencer,C.C.A., Pointon,J.J., Su,Z., Harvey,D., Kochan,G., Oppermann,U., Dilthey,A., Pirinen,M., Stone,M.A. *et al.* (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.*, **43**, 761–767.
5. Liu,Y., Xu,H., Chen,S., Chen,X., Zhang,Z., Zhu,Z., Qin,X., Hu,L., Zhu,J., Zhao,G.-P. *et al.* (2011) Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS Genet.*, **7**, e1001338.
6. Strange,A., Capon,F., Spencer,C.C., Knight,J., Weale,M.E., Allen,M.H., Barton,A., Band,G., Bellenguez,C., Bergboer,J.G. *et al.* (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.*, **42**, 985–990.
7. Wei,W., Hemani,G., Hicks,A.A., Vitart,V., Cabrera-Cardenas,C., Navarro,P., Huffman,J., Hayward,C., Knott,S.A., Rudan,I. *et al.* (2011) Characterisation of genome-wide association epistasis signals for serum uric acid in human population isolates. *PLoS One*, **6**, e23836.
8. Cordell,H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.
9. Hemani,G., Theocharidis,A., Wei,W. and Haley,C. (2011) EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, **27**, 1462–1465.
10. Kam-Thong,T., Czamara,D., Tsuda,K., Borgwardt,K., Lewis,C.M., Erhardt-Lehmann,A., Hemmer,B., Rieckmann,P., Daake,M., Weber,F. *et al.* (2011) EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *Eur. J. Hum. Genet.*, **19**, 465–471.
11. Schupbach,T., Xenarios,I., Bergmann,S. and Kapur,K. (2010) FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, **26**, 1468–1469.
12. Wan,X., Yang,C., Yang,Q., Xue,H., Fan,X., Tang,N.L.S. and Yu,W. (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am.J. Hum. Genet.*, **87**, 325–340.
13. Yung,L.S., Yang,C., Wan,X. and Yu,W. (2011) GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics*, **27**, 1309–1310.
14. Gauderman,W.J. (2002) Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.*, **155**, 478–484.
15. Thompson,J.R., Attia,J. and Minelli,C. (2011) The meta-analysis of genome-wide association studies. *Brief. Bioinform.*, **12**, 259–269.
16. Wei,W., Hemani,G., Gyenesei,A., Vitart,V., Navarro,P., Hayward,C., Cabrera-Cardenas,C., Huffman,J., Knott,S.A., Hicks,A.A. *et al.* (2012) Genome-wide analysis of epistasis in body mass index using multiple human populations. *Eur. J. Hum. Genet.*, February 15 (doi:10.1038/ejhg.2012.17; epub ahead of print).
17. Lam,A.C., Powell,J., Wei,W.H., de Koning,D.J. and Haley,C.S. (2009) A combined strategy for quantitative trait loci detection by genome-wide association. *BMC Proc.*, **3(Suppl. 1)**, S6.
18. Wei,W.H., Knott,S., Haley,C.S. and de Koning,D.J. (2010) Controlling false positives in the mapping of epistatic QTL. *Heredity*, **104**, 401–409.
19. Sabatti,C., Service,S.K., Hartikainen,A.L., Pouta,A., Ripatti,S., Brodsky,J., Jones,C.G., Zaitlen,N.A., Varilo,T., Kaakinen,M. *et al.* (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.*, **41**, 35–46.
20. Wang,K., Li,M. and Hakonarson,H. (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 843–854.
21. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.